New Investigator Tools for Finding Unique and Common Components in Multiple Samples with Comprehensive Two-Dimensional Chromatography

Qingping Tao;[1] Stephen E. Reichenbach;[1,2] Chase Heble;[1] Zhanpin Wu;[3]

[1] GC Image, LLC, Lincoln NE, USA

[2] University of Nebraska, Lincoln NE, USA

[3] Zoex Corporation, Houston TX, USA

Comprehensive two-dimensional chromatography is a powerful technique for highly effective chemical separations of complex mixtures and increasingly is used for cross-sample analyses such as sample classification and biomarker discovery. These techniques, such as GCxGC and LCxLC, produce large data sets that are rich with information, but highly complex, and that require automated processing with robust methods. An important challenge is to select a few markers that can be used effectively for clustering and classifying multiple samples. A newly developed workflow and associated tools allow analysts to detect common and unique compounds across many samples with specialised detection and identification constraints that use chromatographic and mass spectral information to distinguish marker compounds. In addition, new visualisation tools for multi-classification methods provide not only metric values, but also instructive predictions as to which features are effective for distinguishing samples.

Introduction

Untargeted cross-sample analyses such as sample classification and biomarker discovery require separating, quantifying, and identifying a large number of compounds in chemically rich samples and then relating the complex compositions across samples and sample classes. Advanced chromatography, mass spectrometry, and statistical data analysis methods can be combined to address this challenge [1]. In particular, separations performed with comprehensive two-dimensional chromatography (such as GCxGC and LCxLC) provide much greater separation capacity and signal-to-noise ratio than traditional one-dimensional chromatography [2,3]. Coupled with high-resolution accurate mass spectrometry, comprehensive two-dimensional chromatography is a powerful analytical solution. However, the large and complex data sets produced also present challenges for data analysis.

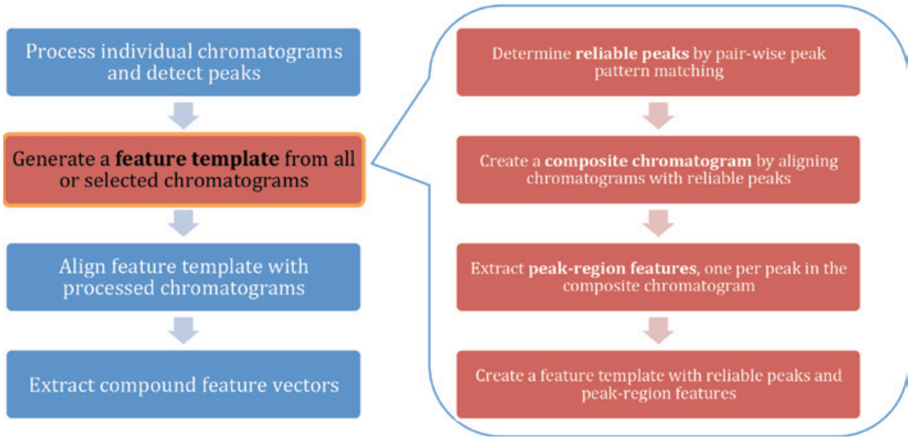
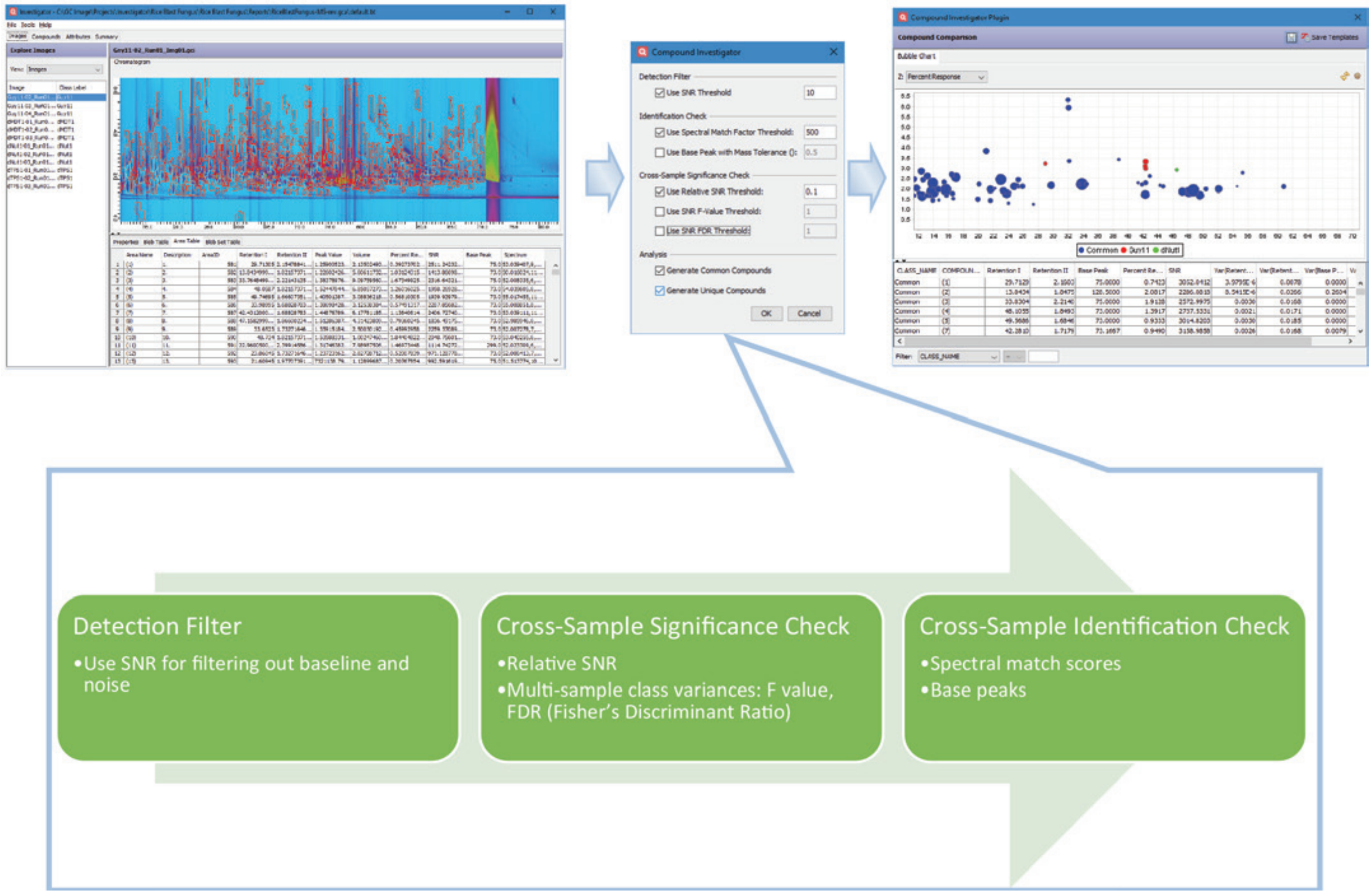

Process individual chromatograms and detect peaks
Generate a **feature template** from all or selected chromatograms
Align feature template with processed chromatograms
Extract compound feature vectors
Determine **reliable peaks** by pair-wise peak pattern matching
Create a **composite chromatogram** by aligning chromatograms with reliable peaks
Extract **peak-region features**, one per peak in the composite chromatogram
Create a feature template with reliable peaks and peak-region features

Figure 1. Automated Investigator Workflow. The workflow detects and aligns all compounds from multiple chromatograms, and extracts feature vectors for untarged cross-sample analyses.

The Investigator™ framework (GC Image, Lincoln NE, USA) developed previously analyses data from multiple samples to extract a feature template that comprehensively captures the pattern of peaks detected in the retention-times plane [4, 5]. Automated feature template extraction, as outlined in Figure 1, is performed by: (1) matching peaks to construct a pattern of alignment peaks that can be reliably matched across chromatograms; (2) aligning and combining chromatograms across samples to create a composite chromatogram; and (3) detecting peak-regions observed in the composite chromatogram. Then, for each sample chromatogram, the extracted feature template is transformed to align with the detected peak pattern and used to generate a set of feature measurements from transformed peak-regions for cross-sample analyses. The approach avoids the typically intractable problem of comprehensive peak matching and can produce feature templates with thousands of features.

The result of the Investigator framework is a feature database with three data dimensions: the chemical features extracted (i.e., peaks and peak-regions); the various attributes measured for each feature, such as retention



Compound Investigator

Detection Filter
☑ Use SNR Threshold 10

Identification Check
☑ Use Spectral Match Factor Threshold: 500
□ Use Base Peak with Mass Tolerance (): 0.5

Cross-Sample Significance Check
☑ Use Relative SNR Threshold: 0.1
□ Use SNR F-Value Threshold: 1
□ Use SNR FDR Threshold: 1

Analysis
☑ Generate Common Compounds
☑ Generate Unique Compounds

OK Cancel

CLASS_NAME	COMPOUN...	Retention I	Retention II	Base Peak	Percent Re...	SNR	Var(Retent...	Var(Retent...	Var(Base P...
Common	(1)	29.7129	2.1600	73.0000	0.7423	3052.8412	3.5795E-6	0.0078	0.0000
Common	(2)	13.8434	1.8475	128.5000	2.0817	2286.8818	8.5415E-6	0.0266	0.2604
Common	(3)	33.8304	2.2140	73.0000	1.9138	2572.9975	0.0036	0.0168	0.0000
Common	(4)	48.1055	1.8493	73.0000	1.3917	2737.5331	0.0021	0.0171	0.0000
Common	(5)	49.5686	1.6846	73.0000	0.9333	3014.8200	0.0030	0.0185	0.0000
Common	(7)	42.2810	1.7179	73.1667	0.9490	3158.9858	0.0026	0.0168	0.0079

Filter: CLASS_NAME

Detection Filter
- Use SNR for filtering out baseline and noise

Cross-Sample Significance Check
- Relative SNR
- Multi-sample class variances: F value, FDR (Fisher's Discriminant Ratio)

Cross-Sample Identification Check
- Spectral match scores
- Base peaks

Figure 2. New Workflow UI and Visualisation. The new workflow uses specialised detection and identification constraints to search for common and unique compounds. The result is displayed in a colour-labelled bubble plot that can be examined by analysts.

times and responses; and the samples and sample classes, for which relative measurements can be used to compare compositions. Such feature databases can be used for chemical fingerprinting, sample classification, chemical monitoring, sample clustering, and biomarker discovery. One important challenge is to develop data analysis and visualisation tools that can help select a few markers that can be used effectively for clustering and classifying multiple samples.

One common problem of marker selection is to detect unexpected compounds that appear in some samples but not others. A new workflow and associated tools are developed to allow analysts to detect common and unique compounds across many samples. This new workflow extends the Investigator framework with specialised detection and identification constraints that use chromatographic and mass spectral information to distinguish targeted compounds. In addition, new visualisation tools for multi-classification methods provide not only metric values, but also instructive predictions as to which features are effective for distinguishing samples. The workflow is demonstrated with two sample sets analysed by GCxGC coupled with quadrupole time-of-flight (Q-TOF) mass spectrometry.

Classical Statistics

Given a feature database extracted by the Investigator framework, classical statistical tools can be used for multiclass analysis to select constituents whose relative presence in samples are statistically related to the classes of the samples. For two sample classes, the Fisher Discriminant Ratio (FDR) is often used. FDR is the ratio of between-group variance to within-group variance [6, 7]. It can be used to assess pairwise class differences for the measure in each peak-region feature:

$$FDR(x_1, x_2) = \frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2)} \, ,$$

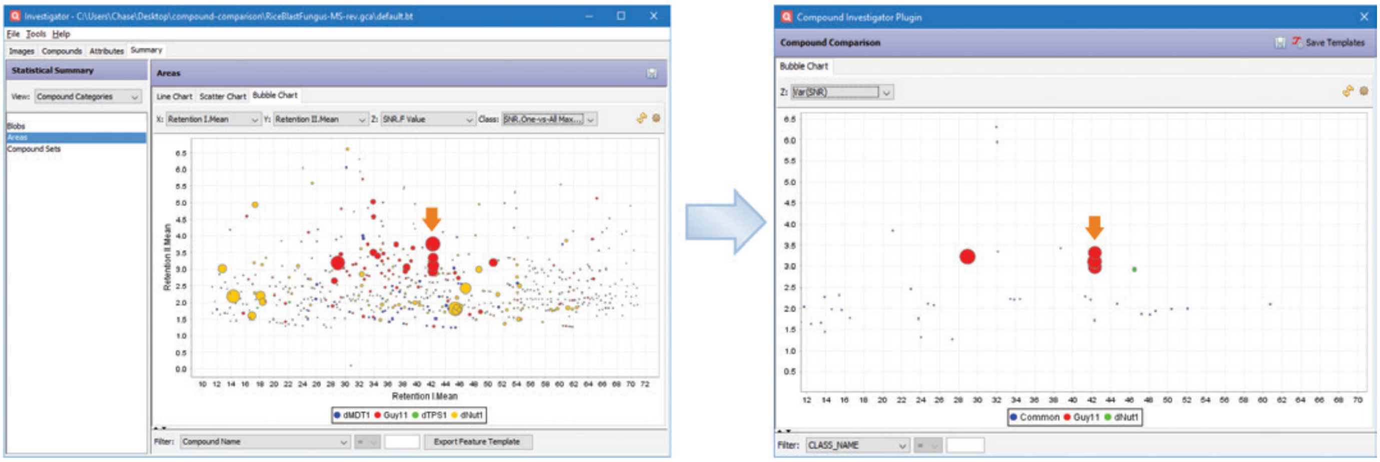
where FDR(x_1, x_2) is the FDR for the sample sets of measured values from two classes, x_1 from Class 1 and x_2 from Class 2; μ_i is the mean of sample values in x_i; and σ_i^2 is the variance of sample values in x_i. For multiple sample classes, the F value is used to assess multiclass differences [7, 8]:

$$F(x_1, \ldots x_K) = \frac{\sum_i N_i(\mu_i-\mu)^2/(K-1)}{\sum_{i,j} N_i\left(x_{i,j}-\mu_i\right)^2/(N-K)} \, ,$$

where K is the number of classes, N_i is the number of sample values in x_i, N is the number of sample values in all classes, μ_i is the mean of sample values in x_i, μ is the mean of all sample values, and $x_{i,j}$ is the jth value in x_i. For k=2 and N_1=N_2, FDR and F value are equal.

A large FDR or F value indicates a large separation of the class means relative to their within-class distributions. The direction of the change is indicated by the difference in means.

Although FDR and F value work well for traditional data classification analysis, they do not always accommodate practical requirements of chromatographic data analysis and chemical marker selection. For example, in practice, it may be expensive to collect multiple samples per class or to acquire multiple chromatographic runs for each sample. In situations with a single chromatographic run per sample class, FDR and F value cannot be computed (because they rely on within-class variance). Also, even if multiple chromatogram runs are available for each sample class, FDR or F value alone may not provide reliable predictions of commonality and uniqueness. For example, a compound feature with a high F value may be simply due to the response differences instead of identity differences across all samples. Thus, in order to detect unique compounds that appear in one sample class but not others or common compounds that appear in all samples, multiple attributes measured such as retention times, responses, and spectral information need to be used to cross-check the identities of chemical features.



Investigator - C:\Users\Chase\Desktop\compound-comparison\RiceBlastFungus-MS-rev.gca\default.bt

View: Compound Categories

X: Retention I.Mean Y: Retention II.Mean Z: SNR.F Value Class: SNR.One-vs-All Max...

Filter: Compound Name

Compound Investigator Plugin

Z: Var(SNR)

Filter: CLASS_NAME

Figure 3. Untargeted Analysis of 4 Types of Rice Blast Fungus. The left bubble plot shows all compound features with F values as bubble sizes. The right bubble plot shows only the compound features that are detected as either unique or common with variances as bubble sizes. Orange arrows point out some compound features that are detected as unique for Guy11.

Methods

A new workflow and associated tools are developed to extend the Investigator framework with specialised detection and identification constraints based on chromatographic and mass spectral information to distinguish targeted compounds. Their purpose is to provide searching and visualisation methods that operate on a feature database to find common and unique compounds across many samples for either multi-sample classes or one-sample classes.

Starting with the Investigator framework, each chromatogram is analysed using a template comprising: (1) a set of peaks that can be reliably recognised across chromatograms, which are used for chromatographic alignment, and (2) a comprehensive set of peak-regions, which are used as features for semi-quantitative sample comparisons. The reliable peaks are determined from the bidirectional pairwise matching of all possible pairs of chromatograms [9]. The peak-region features are delineated by peak detection in the composite chromatogram created by aligning and summing all chromatograms [5]. For the analysis of each chromatogram, the template is aligned using the reliable peaks, then each peak-region is regarded as a compound feature. The problem of recognising the same compound feature in each chromatogram is automatically and implicitly resolved because measures are taken in the same peak-region aligned for each chromatogram.

Afterwards, both chromatographic and spectral information of compound features are extracted from chromatograms and deposited to a feature database. The chromatographic attributes extracted include retention times, signal-noise ratio (SNR), and the total intensity count (TIC) in each peak-region. The TIC value provides a relative measure for a compound feature in a chromatogram. In order to normalise across chromatograms, the TIC measures are normalised by the total TIC values for all peak-regions in the same chromatogram to give a measure of percent-response. The spectral information extracted includes the spectrum of each compound feature and its base peak, which can be used to confirm feature correspondences based on spectral similarity. Then, a Compound-Sample-Class hierarchical association index (HAI) is built and pruned by applying specified criteria that can give analytically useful information on compound deviations between samples as shown in Figure 2.

The pruning process on the HAI uses the following three general filters:

- Detection Filter: SNR is used to filter out compound features from peak-regions that contain only background signals.
- Cross-Sample Significance Check: For multi-sample classes, variance-based statistics (i.e., FDR or F value) are used to select compound features. Low variances indicate commonality and high variances indicate uniqueness. For single-sample classes, samples are selected by a threshold applied to relative measures normalised by the maximum value across samples. If all samples are selected, the corresponding compound feature may be common. If only one sample is selected, the compound feature may be unique.
- Cross-Sample Identification Check: The spectra of the same compound feature are compared across samples by match scores and base peaks. A compound unique to a sample should have low match scores when compared with other samples. A compound common among all samples should have high match scores across all samples.

The pruning result is visualised with a color-labeled bubble plot. Each bubble represents a common or unique compound feature found. The colour of the bubble indicates the class for which the compound is detected. The size of the bubble can be set to indicate its significance, for example, SNR for single-sample classes or F value for multi-sample classes. All bubbles are placed based on their retention times. The resulting bubble plot provides not only metric values, but also instructive predictions as to which features are effective for distinguishing samples as demonstrated in the following results.

Experiments

Two example analyses are presented here to demonstrate the effectiveness of the new workflow. The data were processed and visualised using a developmental release of GC Image GCxGC-HRMS Edition Software (Version 2.7, GC Image, Lincoln NE, USA).

Multi-Sample Class Example: Rice Blast Fungus

The first example analysed data from 4 types of rice blast fungus (Magnaporthe oryzae) including the wild-type (wt) Guy11 strain and mutant strains resulting from

Table 1 GCxGC-QTOFMS Instrument Conditions - Rice Blast Fungus

Parameter	Setting
GC Conditions	
Primary Column	HP-5MS UI, 15 m × 0.25 mm × 0.25 µm
Secondary Column	SGE BPX-50, 3.25 m × 0.1 mm × 0.1 µm
Split Ratio	15:1
Split Inlet Temperature	280°C
Oven Temperature Program	60°C to 310°C at 3°C/min
Carrier Gas Flow	1.2 mL/min
Modulation Conditions	
Modulator	Zoex ZX2
Modulation Period	6.8 seconds
Cold Jet Flow	13 L/min
Hot Jet Temperature	375°C
MS Conditions	
Transfer Line Temperature	310°C
Ionisation Mode	EI
Data Acquisition Rate	50 Hz

the deletion of genes encoding a nitrogen regulator (Δnut1), a carbon regulator (Δmdt1), and an integrator of carbon and nitrogen metabolism (Δtps1) [10, 11]. Three samples were collected for each of the four classes (wt, Δnut1, Δmdt1, Δtps1). Mycelial tissue samples were collected, lyophilised, and ground in liquid nitrogen. The metabolites were extracted using a mixture of methanol:chloroform:water (1:2.5:1, v/v/v). The extracts were dried under vacuum and derivatised by methoxination followed by silylation with MSTFA + 1% TMCS. The 12 samples were analysed using a GCxGC-QTOFMS system. The GCxGC system (with Model 7890B GC, Agilent Technologies, Santa Clara CA, USA) employed a loop thermal modulator (Model ZX2, Zoex Corporation, Houston TX, USA). The QTOFMS system (7200 Series GC/Q-TOF MS, Agilent Technologies, Inc) acquired high resolution mass spectra of the secondary column effluent at a rate of 50 spectra per second. The instrument conditions are summarised in Table 1.

The Investigator framework extracted 159 reliable peaks used for alignment and 572 peak-regions used to create a feature template. The following criteria were used to search:

- Detection Filter: SNR > 10,
- Cross-Sample Significance Check: SNR F Value Threshold = 5,
- Cross-Sample Identification Check: Spectral Match Factor Threshold = 500.

From the total of 572 features, 35 features were found as common and 5 features were found as unique for two fungus types, as shown in Figure 3. On the left, a bubble plot shows all features with F values as bubble sizes. Each feature is assigned with a colour for the class that has the largest FDR value computed with the one-vs-all strategy [12]. Features with a large FDR or multi-class F value can be regarded as potential biomarkers of metabolomic differences. On the right, a bubble plot shows common and unique features with F values as bubble sizes. Each unique feature is assigned the class label of the sample that it belongs to after pruning by the above criteria. Clearly, not all potential markers are also unique makers for a specific class. The most promising markers can be examined more closely. Figure 4 shows one of the distinctive features found for wt samples.

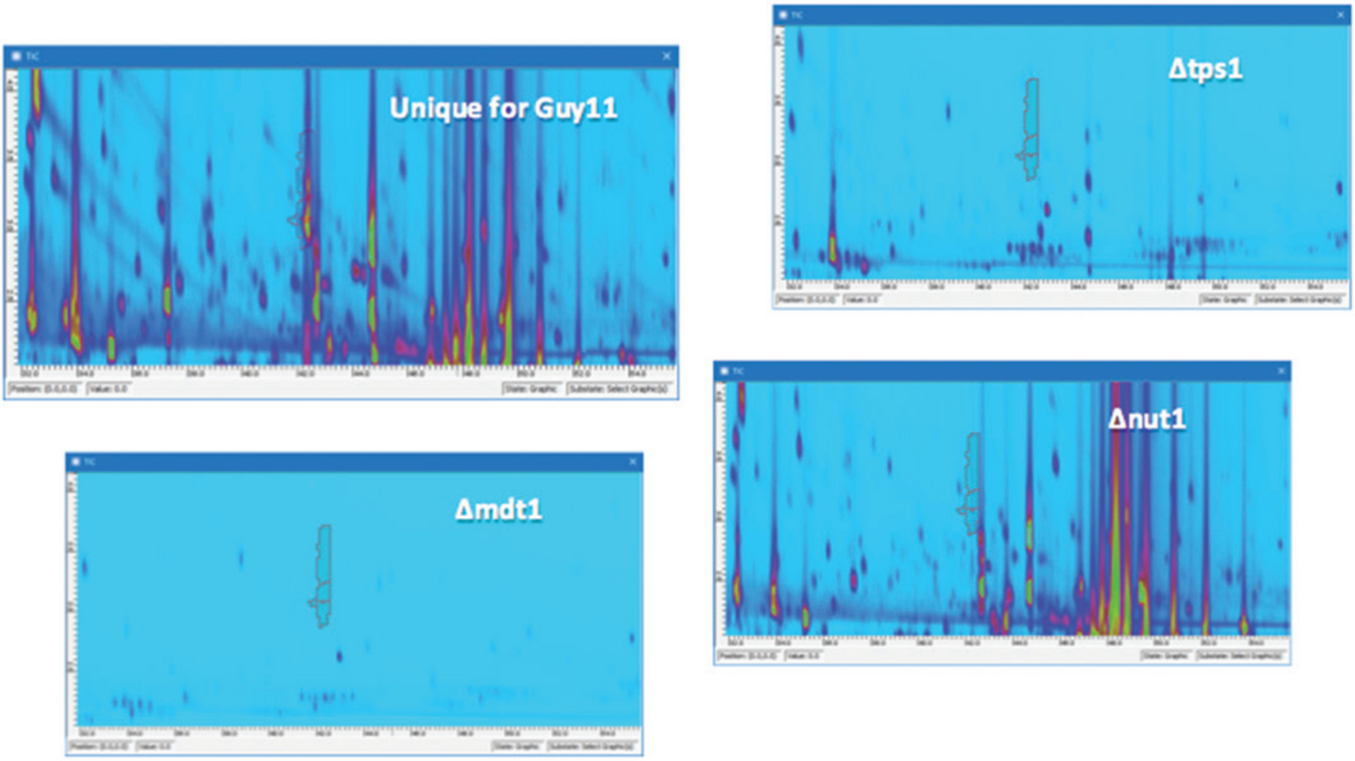


Figure 4. Comparative Results: wt(Guy11) vs. Others. The regions with red outline are compound features found only in wt(Guy11) samples.

Table 2 GCxGC-QTOFMS Instrument Conditions - Essential Oil

Parameter	Setting
GC Conditions	
Primary Column	DB-1, 30 m × 0.25 mm × 0.25 µm
Secondary Column	DB-17, 1.5 m × 0.1 mm × 0.1 µm
Split Ratio	10:1
Split Inlet Temperature	265°C
Oven Temperature Program	45°C to 280°C at 2.2°C/min
Carrier Gas Flow	1 mL/min
Modulation Conditions	
Modulator	Zoex ZX2
Modulation Period	4.5 seconds
Cold Jet Flow	18 L/min
Hot Jet Temperature	170°C to 375°C at 2.5°C/min
MS Conditions	
Transfer Line Temperature	280°C
Ionisation Mode	EI
Data Acquisition Rate	50 Hz

One-Sample Class Example: Essential Oils

The second example analysed data of 10 essential oils including Cardamom, Clove Bud, Coriander, Fennel, Ginger Oil, Juniper Berry, Lavender, Nutmeg, Peppermint, and Turpentine [13]. Only one sample was collected for each type of essential oil. The 10 samples were analysed using the GCxGC-QTOFMS system with Agilent 7890A GC/ Zoex ZX2 thermal modulation system coupled with Agilent 7200 Q-TOF. Samples were directly injected. The instrument conditions are summarised in Table 2.

The Investigator framework extracted 35 reliable peaks used for alignment and 1352 peak-regions used to create a feature template. The following criteria were used to search:

- Detection Filter: SNR > 10,
- Cross-Sample Significance Check: Relative SNR Threshold = 0.1,
- Cross-Sample Identification Check: Spectral Match Factor Threshold = 500.

There are 12 common features and 319 unique features found from total 1352 features as shown in Figure 5. On the left, the composite chromatogram is overlaid with all extracted features indicated by purple rectangles. On the right, a bubble plot shows common and unique features with average percent response as bubble sizes and class labels determined by the above criteria. Figure 6 shows one of the distinctive features for Lavender. In the chromatogram of Juniper Berry, this feature peak-region is a background region; and, in the chromatogram of Ginger Oil, it contains another compound with a different spectrum. Without the cross-sample identity check, this feature would not be found as a unique compound for Lavender.

Conclusion

Classical statistical tools are useful but not sufficient for real-world cross-sample data analysis. The new workflow and associated tools described above combine classical statistical tools with advanced data processing, filtering, and visualisation in order to detect common and unique compounds across multiple samples. The workflow was demonstrated with two typical untargeted analysis cases with GCxGC-QTOFMS data. The same workflow can be used to analyse multiple classes of samples with any comprehensive two-dimensional chromatography technique.

References

1. Alonso, Arnald, Sara Marsal, and Antonio Julià. "Analytical Methods in Untargeted Metabolomics: State of the Art in 2015." Frontiers in Bioengineering and Biotechnology 3 : 23. 2015.
2. Tranchida, P. Q., Franchina, F. A., Dugo, P. and Mondello, L. "Comprehensive two-dimensional gas chromatography-mass spectrometry: Recent evolution and current trends." Mass Spec Rev, 35: 524–534. 2016.
3. Isabelle François, Koen Sandra, Pat Sandra, "Comprehensive liquid chromatography: Fundamental aspects and practical considerations—A review", In Analytica Chimica Acta, Volume 641, Issues 1–2, Pages 14-31, 2009.
4. S. Reichenbach, X. Tian, Q. Tao, E. Ledford, Z. Wu, O. Fiehn. "Informatics for Cross-Sample Analysis with Comprehensive Two-Dimensional Gas Chromatography and High-Resolution Mass Spectrometry (GCxGC-HRMS)." Talanta, 83(4):1279-1288, 2011.
5. S. Reichenbach, X. Tian, C. Cordero, Q. Tao. "Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography." Journal of Chromatography A, 1226:140-148, 2012.

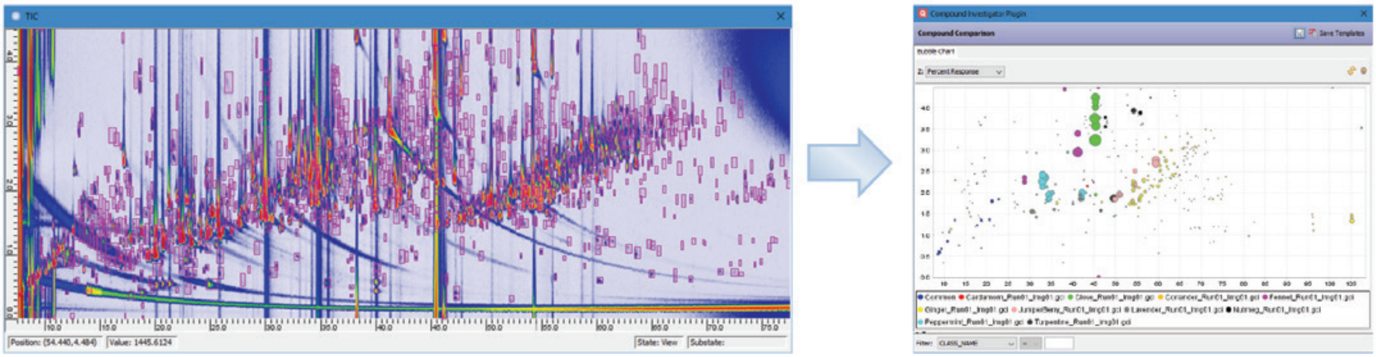


Figure 5. Untargeted Analysis of 10 Essential Oils. The new workflow is able to find potential common and unique compound features (right) from all features of all samples (left).

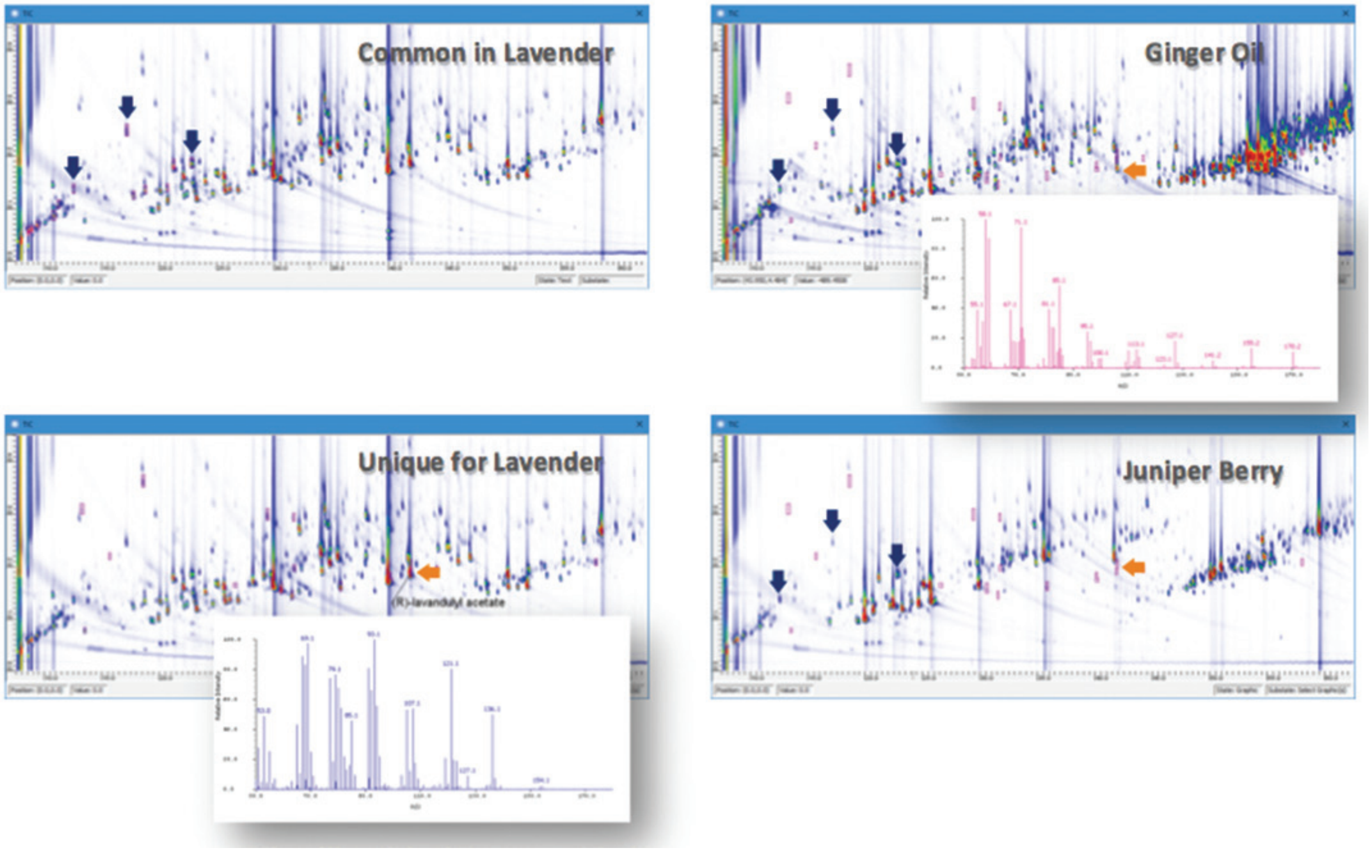


Figure 6. Comparative Results: Lavender vs. Common and Others. Dark blue arrows point out three of the common compound features that exist in all samples. Orange arrows point out one of the unique compound features for lavender. Spectra at the same location also confirm its uniqueness across samples.

6. R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol. 7, no. 2, pp. 179-188, 1936.

7. R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, New York NY: Wiley, 1973.

8. R. C. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," Journal of the Royal Statistical Society, vol. 10, no. 2, pp. 159-203, 1948.

9. S. E. Reichenbach, X. Tian, A. A. Boateng, C. A. Mullen, C. Cordero and Q. Tao, "Reliable Peak Selection for Multisample Analysis with Comprehensive Two-Dimensional Chromatography," Analytical Chemistry, vol. 85, p. 4974–4981, 2013.

10. S. Aronova, W.C. Ledford, M. Marroquin-Guzman, R.A. Wilson, Q. Tao, S.E. Reichenbach, Z. Wu, E.B. Ledford, J. Gushue, and H. Prest. "Untargeted Metabolomics Study of the Plant-Pathogenic Fungus Magnaporthe Oryzae by GCxGCxQTOFMS." GCxGC Symposium, Riva del Garda, IT, May 2014.

11. W.C. Ledford, M. Marroquin-Guzman, R.A. Wilson, E.B. Ledford, Z. Wu, Q. Tao, S.E. Reichenbach, S. Aronova. "Using GCxGC and the Agilent 7200 GC/Q-TOF for an Untargeted Metabolomics Study of the Fungal Rice Pathogen Magnaporthe oryzae." Agilent Application Note 5991-6518EN, 2016.

12. C.M., Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

13. E. Ledford, Z. Wu, S. Nieto, S. Reichenbach, and Q. Tao. "GCxGCxQ-TOF-MS Survey of Essential Oils." ASMS Conference on Mass Spectrometry and Allied Topics, San Antonia TX, June 2016.